Published in final edited form as:

Biopolymers. 2013 November ; 100(6): 780–789. doi:10.1002/bip.22305.

A serendipitous survey of prediction algorithms for amyloidogenicity

Bartholomew P. Roland, Ravindra Kodali, Rakesh Mishra, and Ronald Wetzel

Department of Structural Biology and Pittsburgh Institute for Neurodegenerative Diseases,
University of Pittsburgh School of Medicine, Biomedical Sciences Tower 3, 3501 Fifth Avenue,
Pittsburgh, PA 15260

SUMMARY

The 17- amino acid N-terminal segment of the Huntingtin protein, htt^{NT}, grows into stable α -helix rich oligomeric aggregates when incubated under physiological conditions. We examined 15 scrambled sequence versions of an htt^{NT} peptide for their stabilities against aggregation in aqueous solution at low micromolar concentration and physiological conditions. Surprisingly, given their derivation from a sequence that readily assembles into highly stable α -helical aggregates that fail to convert into β -structure, we found that three of these scrambled peptides rapidly grow into amyloid-like fibrils, while two others also develop amyloid somewhat more slowly. The other 10 scrambled peptides do not detectably form any aggregates after 100 hrs incubation under these conditions. We then analyzed these sequences using four previously described algorithms for predicting the tendencies of peptides to grow into amyloid or other β -aggregates. We found that these algorithms – Zygggregator, Tango, Waltz and Zipper – varied greatly in the number of sequences predicted to be amyloidogenic and in their abilities to correctly identify the amyloid forming members of scrambled peptide collection. The results are discussed in the context of a review of the sequence and structural factors currently thought to be important in determining amyloid formation kinetics and thermodynamics.

Keywords

thermodynamics; kinetics; oligomer; nucleation; Zipper; Zygggregator; Tango; Waltz; amyloid; aggregate; β -structure; huntingtin; polyglutamine

Protein aggregation and/or amyloid formation is an important intrinsic constraint on protein sequence evolution^{1–4}, the focus of a variety of cellular housekeeping operations⁵, a significant challenge in biotechnology^{6–9} and a component of the molecular mechanisms of a number of important human neurodegenerative diseases^{10,11}. Significant effort has been expended in developing an understanding of the molecular determinants of protein aggregation, and in particular in devising algorithms that might be predictive for the propensity of amino acid sequences to aggregate^{12–20}. The fundamental approaches taken in both the development of these algorithms and their validation differ considerably. Empirical assessments of these algorithms have been limited, however, and, where attempted, have in some cases been restricted to amino acid compositions that are highly favorable in promoting aggregation.

Recently, we defined several contributions of the short, 17-amino acid peptide segment called htt^{NT}, which is located at the N-terminus of the Huntington's disease protein huntingtin (htt), to spontaneous amyloid formation by N-terminal fragments of htt such as htt exon1^{21–26}. For example, we found that the htt^{NT} segment, as part of exon1-like fragments, greatly enhances nucleation of amyloid growth by the adjacent polyglutamine (polyQ) segment, without itself becoming incorporated into amyloid^{21–23}. We also found

that the htt^{NT} sequence alone, as a separate peptide, is an effective, if transient, inhibitor of amyloid nucleation by exon1-like fragments²⁵. As part of an effort to determine the molecular basis for this inhibition, we generated a series of scrambled variants of the htt^{NT}Q sequence²⁵. Most of these sequences were well-behaved under our experimental conditions and could be evaluated as inhibitors in our experiments. Some of them could not be evaluated, however, because they themselves undergo efficient spontaneous growth of β -aggregates under our experimental conditions. This was especially intriguing because the WT htt^{NT} sequence by itself aggregates very slowly and only at high concentrations, and then only to form α -helix rich oligomers^{22,23}. It struck us that the behavior of this set of scrambled peptides might serve as a good test for how well previously described amyloid prediction algorithms can predict facile amyloid formation in a novel series of sequences not designed for that purpose.

In this paper we report the spontaneous aggregation tendencies of the full set of scrambled sequences and evaluate how well a number of recently described predictive algorithms are able to distinguish aggregating peptides from peptides that remain soluble under physiological conditions. The results show that the available algorithms are only modestly successful at predicting amyloid propensity under conditions approaching physiological. In fact, it is not clear that any of these algorithms is capable of evaluating kinetics constraints on amyloid formation, even among sequences that might be capable of achieving snug steric zipper²⁷, amyloid-like packing arrangements.

RESULTS

The htt^{NT} sequence MATLEKLMKAFESLKSF is a highly conserved 17-amino acid segment at the very N-terminus of the ~3500 amino acid huntingtin protein responsible for the expanded polyQ repeat malady Huntington's disease²⁸. This N-terminal segment plays significant roles in the normal and abnormal biology of the huntingtin protein. The presence of this peptide segment in *cis* dramatically enhances polyQ amyloid formation²¹, and in *trans* is a good inhibitor of htt exon1 aggregation²⁵. In a transgenic mouse model, only two amino acid replacements within the htt^{NT} segment of a full-length htt protein containing a long expanded polyQ repeat abrogates neuronal aggregate accumulation, HD symptoms, and early death^{24,29}. In isolation in solution, this sequence exists in an equilibrium between a disordered monomer²¹ and an α -helix rich tetramer²³, and upon incubation in PBS at 37 °C it undergoes a slow, concentration dependent conversion into an α -helix rich, sedimentable oligomer^{21–23}. Even after nucleation of polyQ amyloid growth within this oligomer²³, the htt^{NT} sequence retains its α -helical structure in the final polyQ-core amyloid fibrils^{22,23}.

To study the sequence and structural constraints on this α -helical assembly and consequent inhibition of amyloid nucleation²⁵, using a random sequence generator we designed a series of 20 scrambled peptides derived from the htt^{NT}Q sequence and obtained small amounts of these peptides in crude state by custom synthesis (Methods). Five of these 20 peptides could not be evaluated due to poor synthetic yields. The remaining 15 sequences (Table 1) were evaluated as possible inhibitors of the amyloid formation of an exon1-like peptide²⁵. While 12 of these sequences retained good solubility over the time frame of the inhibition experiments, three aggregated rapidly, even when incubated alone, and were not pursued further as potential inhibitors.

Aggregation by scrambled htt^{NT}Q peptides

The results of a survey of the aggregation propensity of these 15 peptides at 6 μ M in pH 7.4 PBS at 37 °C are displayed in Figure 1, which shows the percentage of the starting monomer remaining in solution at various times, as determined by HPLC analysis after centrifugation to remove aggregates (Methods). Consistent with previous reports^{21,23}, the WT htt^{NT}Q

peptide aggregates slightly under these conditions. Most of the scrambled peptides incubated under identical conditions exhibited, within error, no aggregation up to 4 days (Fig. 1). However, three scrambled sequence peptides, SP10, SP14, and SP15, aggregated significantly over the first 10–20 hrs, and another two peptides, SP11 and SP13, aggregated much more slowly but after 3–4 days had aggregated about 30–40% to completion (Fig. 1).

To determine the type of aggregates formed by these different peptides, we examined the aggregated products by EM and FTIR, in some cases after scaling up the reactions to obtain sufficient material for analysis (Methods). As previously reported, negative stained EMs of the aggregates produced by incubation of htt^{NT} peptides with short or missing polyQ segments are amorphous in appearance (Fig. 2 A) and exhibit FTIR spectra consisting in large part of α -helix (Fig. 3 A; Fig. 4), even after over 1,000 hrs incubation at low mM concentrations²³. In contrast, the EM images of the scrambled peptide aggregates exhibit various filamentous morphologies suggesting amyloid structures (Fig. 2 B–F), and the amyloid-like character of these aggregates was further supported by the presence of a strong β -sheet band in the 1622 – 1626 cm⁻¹ range in the FTIR (Fig. 3 B–F; Fig. 4).

We conducted additional experiments to further probe the relationship between sequence and amyloidogenicity in this series. In an unrelated examination of alanine substitutions in htt^{NT}-derived fragments (R. Mishra and R. Wetzel, unpublished) we observed that a Lys to Ala point mutation at position 6 in an htt^{NT}Q₃ sequence, when incubated at 37 °C at 20–30 μ M in PBS, aggregates faster than the WT sequence (Fig. 5 A) and, in contrast to the WT sequence, forms β -sheet-rich (Fig. 5 B), amyloid-like (Fig. 2 G) fibrils. Since the same FTIR spectrum is observed at both a relatively early and a late incubation time (Fig. 5 B), it is likely the small amount of α -helix seen in the FTIR is probably a component of the final aggregates and not some residual oligomeric intermediate. These data are in contrast to the stable formation of α -helix rich oligomeric aggregates when the WT htt^{NT}Q₃ sequence is incubated under the same conditions.

Analysis of scrambled sequences with aggregation predictor algorithms

We took advantage of the availability of the above aggregation data to evaluate the ability of several available algorithms representing highly diverse approaches to predicting amyloidogenic primary sequences. One of these algorithms, Zyggregator¹⁶, computes the aggregation tendency of local primary sequence based on hydrophobicity, charge, the balance between β - and α -potential, and the degree to which amino acids are arranged in an alternating pattern optimal for β -sheet formation. We also evaluated Tango¹⁵, a statistical mechanics based algorithm for predicting any kind of β -structure-based aggregation, that takes into account hydrophobicity, charge, and H-bonding contributions, and the balance in propensities for β -structure, β -turns, and α -structure. We also evaluated Waltz, an extension of Tango that was vetted against the positional preferences for different amino acids in a training set consisting of a large number of diverse, experimentally validated amyloid-forming and non-amyloid forming peptide sequences²⁰. Finally, we evaluated Zipper¹⁹, an algorithm based on 3D-profiling of test sequences against structural models from atomic resolution X-ray analysis of peptide crystals featuring steric zipper packing arrangements of the kind expected to be found in amyloid fibrils²⁷.

Generally, the developers of the Zyggregator algorithm consider peptide sequences to be amyloidogenic if the Zagg score over a patch of residues is, per residue, greater than 1.0³⁰. Five of the 16 htt^{NT}Q sequences contained segments that scored greater than 1.0 (Table 1). Slowly aggregating SP11 and SP13 both received Zagg scores much greater than 1.0. However, none of the three rapidly aggregating scrambled htt^{NT}Q peptides scored higher than 0.83. Of the other three peptides (SP3, SP8, and SP9) receiving scores greater than 1.0, none was observed to aggregate under our experimental conditions up to 90 hrs.

Zygggregator also failed to predict the transformation of the non-amyloid WT htt^{NTQ} to an efficient amyloid forming peptide by the K6A mutation, since the Zagg score moved only from 0.331 to 0.375 (Table 1) in response to this mutation.

Analysis by the Tango software with test conditions set to match our *in vitro* conditions of pH 7.4, ionic strength of 0.14 M, and 37 °C, produced high β -aggregation scores for five of the sequences. The developers of the software consider that a sequence with five contiguous residues of >5% aggregation propensity for each residue is quite likely to make β -aggregates under their standard conditions (which include 1 mM peptide concentrations)¹⁵. The values in Table 1 are the sums over the individual residue scores, so values over 25 are considered by Tango to be likely to aggregate. Table 1 shows that of the scrambled peptides that aggregate under our conditions, only one – the slowly aggregating SP11 – received a score predictive of aggregation. All three peptides that grow into amyloid fibrils within four days at 6 μ M received very low Tango β -aggregation scores, while none of the other high scoring peptides was found to aggregate (into amyloid or anything else). The Tango β -aggregation scores for both WT and K6A htt^{NTQ} are 0. Both Zygggregator and Tango predict five peptides to be susceptible to β -aggregation, and three of these predictions of aggregation propensity – for SP8, SP9, and SP11 – are shared by the two algorithms.

The Waltz algorithm produced the most stringent amyloid predictions of the four programs tested. Only two sequences were predicted by Waltz to be amyloidogenic, SP8 and SP14. One of these, SP14, rapidly grows into amyloid fibrils, while the other, SP8, exhibits no aggregation after 90 hrs under our 6 μ M test conditions. Waltz correctly predicted no amyloid growth for WT and nine other scrambled htt^{NTQ} sequences, but also failed to predict aggregation by four other peptides that do spontaneously aggregate, partially to completion, within four days. Like the other algorithms tested, Waltz failed to detect the amyloidogenic nature of the K6A mutation. Interestingly, non-amyloid forming SP8 is strongly predicted to be aggregation prone by Zagggregator, Tango, and Waltz.

In stark contrast with Waltz, the 3D profiling algorithm Zipper predicted potential amyloid motifs in most of the peptides in the collection. Using the recommended benchmark of a Rosetta energy of -23 or below, and a shape complementarity score of 0.7 or better¹⁹, 13 of the scrambled htt^{NTQ} peptides were found to be of 'high fibrillation potential' (HP). Interestingly, the two peptides that failed to achieve HP scores, SP10 and SP14, turned out to be two of the three rapid aggregators in our kinetics analysis (Fig. 1). In addition, SP8, which did not aggregate in our conditions but which was predicted to be amyloidogenic by Zagggregator, Tango, and Waltz, also scored a highly favorable Rosetta energy of -25.6 in Zipper, the most stable energy for a steric zipper motif found for any of the peptides analyzed.

Given the general agreement among the four algorithms that SP8 should be capable of making amyloid, we obtained this peptide in quantity and incubated it at ~ 0.5 mM concentration. At this high concentration, SP8 formed visible aggregates immediately, and on examination these aggregates exhibited EM morphology (Fig. 2 H) and an FTIR spectrum (Fig. 3 H) consistent with amyloid fibrils. Thus, the algorithms correctly predicted the ability of this sequence to engage the amyloid motif structure, but uniformly failed to give a realistic appraisal of its ability to do so at the relatively low concentrations relevant to most *in vivo* situations.

It is worth considering whether greater accuracy of prediction for the scrambled htt^{NTQ} peptides might be achieved by using some of the algorithms in combination. For example, by demanding qualifying scores for both Zagggregator and Zipper, five of the 15 scrambled htt^{NTQ} peptides are predicted to be amyloidogenic, of which two – SP11 and SP13 – do

aggregate at the 6 μ M conditions. A third, SP14 was barely missed due to a slightly low Zagg score. However, each of the two other rapidly aggregating peptides, SP10 and SP15, were actually missed by both algorithms.

DISCUSSION

Mechanistic components of amyloidogenicity

As for any chemical reaction, the tendency of a particular protein or peptide to make amyloid fibrils will depend on both the thermodynamic and kinetic favorability of the conversion. Each of these factors has its own complexities raising obstacles for accurate predictions of amyloidogenicity.

The thermodynamic favorability is simply the difference between the Gibbs free energy of the monomer in solution and that of the final reaction mixture containing the amyloid fibril product plus any unreacted monomer or other aggregates at equilibrium. Thermodynamic stability must be considered in the context of the concentration of the amyloidogenic sequence; if this is lower than the critical concentration associated with aggregate stability³¹, there is no driving force for aggregate formation, no matter how favorable the kinetics might become at higher concentrations. One consequence of this is that obtaining relative aggregation tendencies at low, cellular concentrations may not be a simple matter of extrapolation from *in vitro* studies that were conducted at much higher concentrations. The importance of thermodynamic favorability is seemingly well-captured by predictive algorithms like Zipper¹⁹ that focuses on how well a particular sequence can be accommodated into a “steric zipper” type packing arrangement. The Waltz algorithm²⁰ also contains some structural information, by requiring test sequences to conform to the positional preferences of each amino acid established for amyloid formation in a series of experimentally tested sequences. These are relative scales, however, allowing comparisons between sequences. It is not clear how the numbers might be converted into real world “solubilities” that can be related to specific peptide concentrations.

Compared to thermodynamic stability, amyloid formation kinetics appears to be much more challenging to model from primary sequence alone. Our understanding of the complexity of amyloid growth kinetics continues to evolve. Spontaneous amyloid formation initially depends on a primary nucleation event. To the extent that the kinetic nucleus resembles the final amyloid structure, we might expect that kinetics of formation, like the thermodynamic stability, will depend on how well the amyloidogenic sequence packs into the fibril core structure. While nucleus stability is expected to contribute to the spontaneous aggregation rate, however, there are clearly many other factors that determine nucleation efficiencies and rates.

For example, nucleation mechanisms can vary dramatically. In some cases, nucleation simply depends on productive assembly of a nucleus from a pool of monomers³², as described in the general thermodynamic model of nucleated polymerization³³. In contrast, in many other cases, the appearance of amyloid fibrils is preceded by the appearance of a variety of oligomeric aggregates, any one of which might serve as either the phase within which nucleation occurs^{23,34} or as an inert phase that potentially could negatively impact nucleation and growth by decreasing the effective monomer concentration. Depending on the nucleation mechanism, the concentration dependence of spontaneous aggregation rates will also vary^{21,32}.

Another little appreciated component of amyloid nucleation is the second order rate constant for elongation. In the thermodynamic model of nucleation, nuclei arise from the ground state monomer in a highly unfavorable reversible equilibrium³³. Without a robust elongation rate,

those nuclei that transiently arise will not successfully engage an elongation cycle, but rather will collapse back into the ground state monomer ensemble. Thus, elongation rate not only impacts the rates of growth of existing fibrils, but also can impact the efficiencies with which nuclei are stabilized and become committed to fibril formation. This critical elongation rate depends on several factors. First, it depends on the concentration of elongation competent monomers in the ensemble. If the monomer is an intrinsically disordered protein³⁵ that divides its time in solution between several conformations, some of which are elongation competent and some of which are not^{36–38}, then the ability of the monomer pool to facilitate nucleation by elongation will be diminished accordingly. In addition, the “dock-and-lock” kinetics of this process involves several microscopic rates and equilibria³⁹, and presumably is potentially as complex and idiosyncratic as any protein folding reaction⁴⁰.

The rates of conversion and stabilities of other stable or metastable species on the folding landscape can play major roles in determining amyloid formation kinetics. For example, a widely acknowledged component of protein aggregation kinetics, including for amyloid formation, is the degree to which the aggregation-prone sequence is buried in rigid, inaccessible higher structure (such as in a globular protein)^{2,41}. Although Zyggregator attempts to build in a factor that takes into account this structural restriction, in the context of fibril growth conditions¹⁸, accurate primary sequence-based algorithms that take tertiary and quaternary structural constraints into account must await the arrival of robust methods for predicting, from primary sequence, not only protein structures but also the stability of those structures. Another complication is the potential to form alternatively aggregated, dead-end structures⁴² whose formation may be in kinetic competition with amyloid formation.

Another way in which primary sequence can enhance primary nucleation of amyloid in a peptide was recently revealed in studies of the role of the subject of this paper, the htt^{NT} sequence, in enhancing polyQ amyloid nucleation²¹. This sequence contributes a huge enhancement to nucleation by independently self-associating into tetramers and possibly higher aggregates²³, leading to a high local concentration of the attached polyQ segments which in turn facilitates nucleation^{21,23}. Since the htt^{NT} sequence provides this great enhancement to the spontaneous amyloid formation rate²¹, as well as the stability of the resulting fibrils²⁴, without itself ever becoming a part of the amyloid core^{22,23}, enhancing sequences in this class would be expected to not necessarily be amenable to predictive algorithms for amyloidogenicity. It has been postulated that segments of amyloidogenic peptides IAPP and A β may play similar, transient roles in nucleation of amyloid formation by these peptides⁴³.

It may also be important to gauge the stability of the final amyloid product if one is to generate an accurate estimate of fibril growth kinetics. This is because of the phenomenon of secondary nucleation, a process that can magnify the initial primary nucleation kinetics by virtue of creating more growth sites in existing fibrils through lateral budding or fragmentation^{33,44}. Fibril stability to shear-induced breakage will be a major factor in the degree to which fragmentation-associated secondary nucleation occurs⁴⁴. Observed initial aggregation rates can be dominated by secondary nucleation, thus obscuring the primary nucleation tendencies of a sequence. Conversely, in terms of the practical context of predictive algorithms, it is not clear at the moment how we can effectively model this potentially dominant secondary nucleation reaction based only on primary sequence data. Fibril stability may depend on both the packing of the core amyloid and on further interactions between filaments²⁴. Taking fibril stability and secondary nucleation into account for predicting amyloidogenicity from primary sequence will depend on as yet

unrealized results such as the ability to predict amyloid structure from primary sequence and the ability to calculate stability to shear-induced fragmentation from such structures.

All of the above factors will presumably play a role in contributing to observed amyloid growth rates in both simple experimental systems and in the biological setting. Many additional factors involving complex cellular pathways for managing aggregate formation⁵ will further complicate the ability to predict aggregation tendencies of proteins in their normal biological settings.

Performance of amyloid prediction algorithms

In this paper we evaluated four algorithms representing significantly different fundamental approaches to predicting amyloid formation. These predictions are compared to a new dataset of aggregation kinetics of a series of scrambled peptide sequences (Fig. 1) using incubation conditions that were informed by our experiences studying pathologically relevant peptide amyloidogenesis. Thus, at the conditions of pH 7.4, 0.14 M salt, and 37 °C, and working with rigorously disaggregated peptides, we previously found that a 3 μ M solution of a Q₃₆ version of the HD-related htt exon1 peptide essentially aggregates to completion within 20 hrs²¹, while a 5 μ M solution of the Alzheimer's amyloid plaque peptide A β ₄₂ aggregates to completion within 100 hrs [S. Chemuru and R. Wetzel, unpublished].

It is clear that none of the algorithms performed very well on this set of scrambled sequence peptides. The algorithms range widely in the stringency with which they find potential amyloid peptides in this collection. Waltz predicts only two of 16, while Zipper predicted 14 of 16. Both Zyggregator and Tango fall in the middle by predicting five of 16 peptides as amyloid formers. Perhaps Zyggregator did a slightly better job overall in this limited survey, in correctly hitting two out of five predicted amyloid forming peptides, and only incorrectly predicting three amyloid formers that did not aggregate under our conditions. There are a number of possible explanations for over-prediction as well as under-prediction in this survey.

One possible contributor to over-prediction of amyloidogenicity within our dataset is the fact that most of the algorithms we tested were originally developed using experimental systems that included peptide concentrations significantly larger than the 6 μ M we used in our experiments. There is mixed evidence in support of this potential factor in our experimental results. Thus, Zipper predicts both WT htt^{NTQ} and SP5 to have a high propensity to aggregate, but neither does so at 6 μ M after four days. On incubation at over 1 mM concentration for over 1,000 hrs, SP5 does make β -rich aggregates. Htt^{NTQ}, however, only accumulates as α -rich aggregates even under these extreme conditions. Another possible contributor to over-prediction is our poor current understanding, as described above, of how peptide sequence contributes to aggregation kinetics, and in particular to nucleation mechanism and efficiency. A particular sequence might be theoretically capable of engaging the amyloid motif to produce very stable fibrils, but if nucleation is inefficient, fibril accumulation will not be observed. Its inclusion of experimental input on aggregation kinetics may explain why Zyggregator performs modestly better than other algorithms in this comparison.

Under-prediction in these results is a little harder to rationalize. The consistent inability of these algorithms to identify scrambled htt^{NTQ} sequences that grow rapidly into amyloid fibrils even at 6 μ M concentrations was indeed quite surprising. Of the three rapidly aggregating peptides in our collection, two were missed by every one of the four algorithms! The third peptide, SP14, was missed by two of the four. The results suggest that the models underpinning these algorithms may be biased in ways that compromise their abilities to

recognize amyloid formers in peptides containing combinations of the amino acid constituents found in htt^{NT}Q. For example, both Zygggregator and Tango were developed by focusing on the aggregation abilities of naturally occurring proteins and peptides, and in particular on known, disease-oriented, amyloid forming polypeptides. Perhaps there is something about the htt^{NT}Q amino acid composition of our peptides that defies the rules extracted from these evolved amyloidogenic sequences. In contrast, Waltz was developed from Tango through the application of an additional filter based on the aggregation abilities of a series of scrambled hexapeptide sequences. Likewise, Zipper depends of 3D profiling from the crystal structure of two particular hexapeptides. Perhaps the restriction to hexapeptide models limits the ability of Waltz and Zipper to perform well on the longer peptide sequences in our dataset. Examination of the rapidly aggregating scrambled htt^{NT}Q peptides provides some support for this hypothesis. For example, an *octapeptide* sequence EKMLLEK from SP14 seems well configured to form duplexes with opposing orientations that are held together by central hydrophobicity and flanking salt bridges. In addition, some amyloidogenic peptide motifs longer than octomers, like A β related peptides, achieve their mature amyloid structures by undergoing reverse turn formation that ultimately involves more than one peptide segment in an amyloid motif that encompasses 25–30 residues^{45,46}. Even in simple polyQ sequences, the ability to make a reverse turn appears to play a critical role in amyloid nucleation^{32,47}. Presumably such possibilities must be somehow taken into account if an algorithm is going to be successful.

The tendency of polypeptide sequences to aggregate is a problem in human disease, normal biology, biotechnology, and molecular evolution. Efforts to develop amino acid sequence based algorithms for assessing the tendencies of peptides to form amyloids or other β -aggregates have been only partially successful. The results of the serendipitous survey described here suggest that there is much to be done before a truly general, discriminating algorithm becomes available. Our results suggest that it will be important to develop methods that accommodate more mixed amino acid sequences and more varied peptide lengths, and that can respond to a range of peptide concentrations. Given the complexity and obscurity of the presumed structural influences on amyloid growth kinetics and thermodynamics, it is not at all obvious that useful, robust predictive algorithms are currently within reach.

METHODS

Materials and general methods

The generation of the scrambled sequences analyzed here is described elsewhere²³. Peptides corresponding to these sequences were obtained in low amounts in crude form by custom synthesis from GenScript, Inc. For follow-up analysis, larger amounts of peptide were obtained by custom synthesis from Keck Biotechnology Center at Yale University (<http://keck.med.yale.edu/ssps/>). Each peptide was purified by reverse-phase HPLC and its concentration in the HPLC pool determined from the A₂₁₅ of the pool and the calculated 215 nm extinction coefficient^{23,48}. In order to conserve limited peptide, disaggregation was conducted using a modified protocol compared to the normal method⁴⁹. In the modification, all the peptides were dissolved in water TFA pH 3.0 and centrifuged at 435,680 g for 1.5 hrs at 4 °C, after which the supernatant was adjusted to PBS conditions and used immediately²³.

Analysis of aggregation reactions and products

Disaggregated peptides were suspended in PBS at 6 μ M and incubated at 37 °C. Aliquots were removed and analyzed by analytical HPLC of centrifugation supernatants as described⁴⁹. For the FTIR analysis, the substantial amounts of aggregate required were obtained by scaling up the aggregation reaction to ~500 μ M of peptide incubated under the

same conditions. The electron microscopy images and FTIR spectra were collected as previously described²³. FTIR spectra were deconvoluted using PeakFit.

Computation

Predictions of amyloid and/or aggregation potential were obtained using web-based calculations following the instructions on the web sites. Zyggregator is available at <http://www.vendruscolo.ch.cam.ac.uk>. Tango is available at <http://tango.embl.de/>. Waltz is available at <http://waltz.switchlab.org/>. Zipper is available at <http://services.mbi.ucla.edu/zipperdb/>.

Acknowledgments

This work was supported by NIH grant R01 GM099718 to RW. We acknowledge James Conway and Alexander Makhov for access to the Structural Biology Department's cryo-EM facility.

REFERENCES

1. Mitraki A, King J. FEBS Lett. 1992; 307:20–25. [PubMed: 1639189]
2. Wetzel R. Trends Biotechnol. 1994; 12:193–198. [PubMed: 7764903]
3. DePristo MA, Weinreich DM, Hartl DL. Nat Rev Genet. 2005; 6:678–687. [PubMed: 16074985]
4. Monsellier E, Chiti F. EMBO Rep. 2007; 8:737–742. [PubMed: 17668004]
5. Tyedmers J, Mogk A, Bukau B. Nat Rev Mol Cell Biol. 2010; 11:777–788. [PubMed: 20944667]
6. De Bernardez Clark E, Schwarz E, Rudolph R. Methods Enzymol. 1999; 309:217–236. [PubMed: 10507027]
7. Fahnert B, Lilie H, Neubauer P. Advances in biochemical engineering/biotechnology. 2004; 89:93–142. [PubMed: 15217157]
8. Simpson RJ. Cold Spring Harbor Protocols. 2010 DOI 10.1101/pdb.top79.
9. Vazquez-Rey M, Lang DA. Biotech Bioeng. 2011; 108:1494–1508.
10. Martin JB. N Engl J Med. 1999; 340:1970–1980. [PubMed: 10379022]
11. Merlini G, Bellotti V. N Engl J Med. 2003; 349:583–596. [PubMed: 12904524]
12. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Nature. 2003; 424:805–808. [PubMed: 12917692]
13. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. J Mol Biol. 2004; 342:345–353. [PubMed: 15313629]
14. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. J Mol Biol. 2004; 341:1317–1326. [PubMed: 15302561]
15. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Nat Biotechnol. 2004; 22:1302–1306. [PubMed: 15361882]
16. Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. J Mol Biol. 2005; 350:379–392. [PubMed: 15925383]
17. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. Proc Natl Acad Sci U S A. 2006; 103:4074–4078. [PubMed: 16537487]
18. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. J Mol Biol. 2008; 380:425–436. [PubMed: 18514226]
19. Goldschmidt L, Teng PK, Riek R, Eisenberg D. Proc Natl Acad Sci U S A. 2010; 107:3487–3492. [PubMed: 20133726]
20. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F. Nature Methods. 2010; 7:237–242. [PubMed: 20154676]
21. Thakur AK, Jayaraman M, Mishra R, Thakur M, Chellgren VM, Byeon IJ, Anjum DH, Kodali R, Creamer TP, Conway JF, Gronenborn AM, Wetzel R. Nat Struct Mol Biol. 2009; 16:380–389. [PubMed: 19270701]

22. Sivanandam VN, Jayaraman M, Hoop CL, Kodali R, Wetzel R, van der Wel PC. *J Am Chem Soc.* 2011; 133:4558–4566. [PubMed: 21381744]
23. Jayaraman M, Kodali R, Sahoo B, Thakur AK, Mayasundari A, Mishra R, Peterson CB, Wetzel R. *J Mol Biol.* 2012; 415:881–899. [PubMed: 22178474]
24. Mishra R, Hoop CL, Kodali R, Sahoo B, van der Wel PC, Wetzel R. *J Mol Biol.* 2012; 424:1–14. [PubMed: 22999956]
25. Mishra R, Jayaraman M, Roland BP, Landrum E, Fullam T, Kodali R, Thakur AK, Arduini I, Wetzel R. *J Mol Biol.* 2012; 415:900–917. [PubMed: 22178478]
26. Wetzel R. *J Mol Biol.* 2012; 421:466–490. [PubMed: 22306404]
27. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D. *Nature.* 2005; 435:773–778. [PubMed: 15944695]
28. Zuccato C, Valenza M, Cattaneo E. *Physiol Rev.* 2010; 90:905–981. [PubMed: 20664076]
29. Gu X, Greiner ER, Mishra R, Kodali R, Osmand A, Finkbeiner S, Steffan JS, Thompson LM, Wetzel R, Yang XW. *Neuron.* 2009; 64:828–840. [PubMed: 20064390]
30. Luheshi LM, Tartaglia GG, Brorsson AC, Pawar AP, Watson IE, Chiti F, Vendruscolo M, Lomas DA, Dobson CM, Crowther DC. *PLoS Biol.* 2007; 5:e290. [PubMed: 17973577]
31. O'Neill B, Shivaprasad S, Kheterpal I, Wetzel R. *Biochem.* 2005; 44:12709–12718. [PubMed: 16171385]
32. Kar K, Jayaraman M, Sahoo B, Kodali R, Wetzel R. *Nat Struct Mol Biol.* 2011; 18:328–336. [PubMed: 21317897]
33. Ferrone F. *Meths Enzymol.* 1999; 309:256–274.
34. Serio TR, Cashikar AG, Kowal AS, Sawicki GJ, Moslehi JJ, Serpell L, Arnsdorf MF, Lindquist SL. *Science.* 2000; 289:1317–1321. [PubMed: 10958771]
35. Dyson HJ. *Q Rev Biophys.* 2011; 44:467–518. [PubMed: 21729349]
36. Bertoncini CW, Jung YS, Fernandez CO, Hoyer W, Griesinger C, Jovin TM, Zweckstetter M. *Proc Natl Acad Sci U S A.* 2005; 102:1430–1435. [PubMed: 15671169]
37. Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, Chellgren BW, Creamer TP, Wetzel R. *J Mol Biol.* 2006; 355:524–535. [PubMed: 16321399]
38. Darnell G, Orgel JP, Pahl R, Meredith SC. *J Mol Biol.* 2007; 374:688–704. [PubMed: 17945257]
39. Cannon MJ, Williams AD, Wetzel R, Myszkowski DG. *Anal Biochem.* 2004; 328:67–75. [PubMed: 15081909]
40. Yanagi K, Sakurai K, Yoshimura Y, Konuma T, Lee YH, Sugase K, Ikegami T, Naiki H, Goto Y. *J Mol Biol.* 2012; 422:390–402. [PubMed: 22683352]
41. Kelly JW. *Curr Opin Struct Biol.* 1998; 8:101–106. [PubMed: 9519302]
42. Laganowsky A, Liu C, Sawaya MR, Whitelegge JP, Park J, Zhao ML, Pensalfini A, Soriaga AB, Landau M, Teng PK, Cascio D, Glabe C, Eisenberg D. *Science.* 2012; 335:1228–1231. [PubMed: 22403391]
43. Abedini A, Raleigh DP. *Protein Eng Des Sel.* 2009; 22:453–459. [PubMed: 19596696]
44. Knowles TP, Waudby CA, Devlin GL, Cohen SI, Aguzzi A, Vendruscolo M, Terentjev EM, Welland ME, Dobson CM. *Science.* 2009; 326:1533–1537. [PubMed: 20007899]
45. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, Delaglio F, Tycko R. *Proc Natl Acad Sci U S A.* 2002; 99:16742–16747. [PubMed: 12481027]
46. Luca S, Yau WM, Leapman R, Tycko R. *Biochem.* 2007; 46:13505–13522. [PubMed: 17979302]
47. Kar K, Hoop CL, Drombosky KW, Baker MA, Kodali R, Arduini I, van der Wel PC, Horne WS, Wetzel R. *J Mol Biol.* 2013; 425:1183–1197. [PubMed: 23353826]
48. Kuipers BJ, Gruppen H. *J Agri Food Chem.* 2007; 55:5445–5451.
49. O'Neill B, Thakur AK, Williams AD, Bhattacharyya AM, Chen S, Thiagarajan G, Wetzel R. *Methods Enzymol.* 2006; 413:34–74. [PubMed: 17046390]

Figure 1.

Aggregation kinetics of WT and scrambled versions of htt^{NTQ}. Loss of monomer from solution over time for peptides incubated at 6 μ M in PBS at 37 °C as determined by an HPLC-based sedimentation assay.

Figure 2.

Electron micrographs of aggregates formed by WT, mutated, and scrambled variations of htt^{NT} , htt^{NTQ} , or $\text{htt}^{\text{NTQ}_3}$. Peptide aggregates are from htt^{NT} (A), SP10 (B), SP14 (C), SP15 (D), SP11 (E), SP13 (F), $\text{htt}^{\text{NTQ}_3}$ (K6A) (G), and SP8 (H). Bar = 50 nm.

Figure 3.

FTIR spectra and curve fitting for isolated aggregates of WT htt^{NT} and scrambled derivatives of htt^{NT}Q. Htt^{NT} (A), SP10 (B), SP14 (C), SP15 (D), SP11 (E), SP13 (F), SP5 (G), and SP8 (H). Vertical dashed lines show normally assigned positions of various protein conformational types or secondary structures, including α -helix (" α "), β -sheet (" β "), random coil (" R "), turns (" T "), and side chain conformations (" S "). The β -sheet band around 1690 cm^{-1} is sometimes assigned to anti-parallel β -sheet.

Figure 4.

Estimated secondary structural contents from deconvoluted FTIR spectra (Fig. 3) for each of the peptide aggregates.

Figure 5.

Effect of a Lys to Ala mutation at position 6 in htt^{NT}Q₃. (A) Kinetics of aggregation of 25 μ M concentrations in PBS at 37 °C of htt^{NT}Q₃ WT (filled circles) and K6A (open circles). (B) FTIR spectra of aggregates of htt^{NT}Q₃ (K6A) isolated at 1 hr reaction and 3500 hr reaction showing a major contribution from β -structure in both aggregates.

Table 1

Scrambled sequences and their observed and predicted aggregation behavior^a.

Name ^b	Sequence	Zagg	Tango	Waltz	Zipper
SP1	TMMKFQLLKSAEEKLFAS	0.932	0	0	-24.1
SP2	MLSLKESAKMFFATKELQ	0.832	18.0	0	-23.3
SP3	KQFTLEMAFLSKALSEMK	1.010	3.9	0	-23.9
SP4	KLAFMLKQAELSSEKTFM	0.550	25.8	0	-25.2
SP5	FAKFASEKLESMTMLQ	0.800	2.2	0	-23.2
SP6	MLTFAEFKSMELKSQLAK	0.678	0	0	-23.6
SP7	ASMFEAQLSKEKKMFTLL	0.756	0	0	-23.4
SP8	ELLAKSEQAKSMLFTFMK	1.401	369	97.99	-25.6
SP9	TKFSSFALLAQKEMLKME	1.027	47.5	0	-24.4
<u>SP10</u>	ETLKMSMFLEAQFKKSAL	0.713	4.8	0	-22.9
<u>SP11</u>	ASSQKKMKEMLAFFTLEL	1.650	337	0	-24.3
SP12	MFSKMAKSLFLLAEKTQE	0.543	257	0	-23.6
<u>SP13</u>	KLELKAASQMEFSFTMKL	1.490	2.6	0	-23.6
<u>SP14</u>	KELKQELFFKASATLMMS	0.828	0.9	96.99	-24.8
<u>SP15</u>	SAFMEKMLLLEKQFKAST	0.421	3.9	0	-21.8
Htt ^{NT} Q _{WT}	MATLEKLMKAFESLKSQ	0.331	0	0	-24.4
Htt ^{NT} Q _{K6A}	MATLEALMKAFESLKSQ	0.375	0	0	-24.4

^a Scores predicting a propensity to make amyloid and/or β -aggregates are indicated in bold italics; see text for a description of how to interpret scores.

^b Peptides aggregating fast at 6 μ M indicated by double underlining, peptides that aggregate but at slower rates indicated by single underlining; see Figure 1 for supporting data.